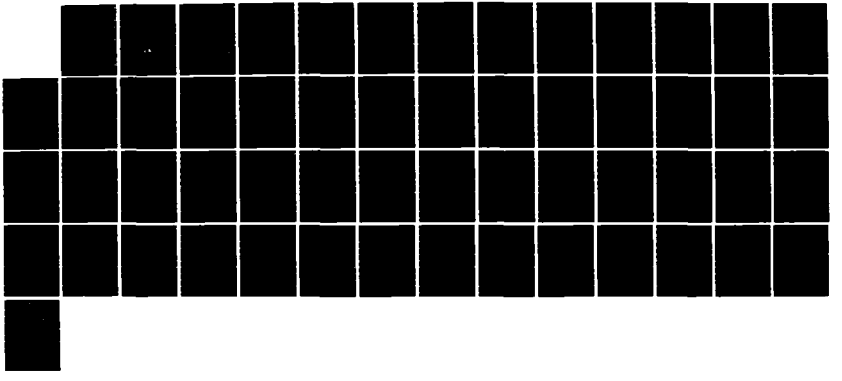
AD-A167 758 THE IMPACT OF PERFORMANCE CONSISTENCY AND PERFORMANCE LEVEL ON ALTERNATIV..(U) MICHIGAN STATE UNIV EAST LANSING DEPT OF PSYCHOLOGY M A YOUTZ ET AL. APR 86 1/1

UNCLASSIFIED TR-86-1 N00014-83-K-0756 F/G 5/10 NL



MICROCOPY CHART

(12)

# MICHIGAN STATE UNIVERSITY

## Industrial/Organizational Psychology and Organizational Behavior

**The Impact of Performance Consistency and Performance Level on Alternative Measures of Rater Accuracy**

by

Margaret A. Youtz and Daniel R. Ilgen

Michigan State University








Michigan State University
East Lansing, Michigan 48824

# The Impact of Performance Consistency and Performance Level on Alternative Measures of Rater Accuracy

by

Margaret A. Youtz and Daniel R. Ilgen

Michigan State University

Prepared for
Office of Naval Research
Organizational Effectiveness Research Programs

Grant No. N00014-83-K-0756
NR170-961



Technical Report No. 86-1
Department of Psychology
Michigan State University

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

| | |
|---|---|
| 1. REPORT NUMBER: 86-1 | 2. GOVT ACCESSION NO.: ADA 167 758 |
| 3. RECIPIENT'S CATALOG NUMBER: | |
| 4. TITLE (and Subtitle): The Impact of Performance Consistency and Performance Level on Alternative Measures of Rater Accuracy | 5. TYPE OF REPORT & PERIOD COVERED: Interim |
| | 6. PERFORMING ORG. REPORT NUMBER: 2011 |
| 7. AUTHOR(s): Margaret A. Youtz and Daniel R. Ilgen | 8. CONTRACT OR GRANT NUMBER(s): N00014-83-K-0756 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology, Michigan State University, East Lansing, MI 48824-1117 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR170-961 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS: Organizational Effectiveness Research Programs, Office of Naval Research (Code 4420E), Arlington, VA 22217 | 12. REPORT DATE: April, 1986 |
| | 13. NUMBER OF PAGES: 49 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) | 15. SECURITY CLASS. (of this report): Unclassified |
| | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

f. Falk

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Rater accuracy, performance consistency, and information processing

This report

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Attempts to understand factors that influence raters' ability to provide accurate ratings have, on occasion, focused on static characteristics of ratees (e.g., race and gender). The present study investigated dynamic characteristics related to the pattern and level of performance displayed by the ratee. In a repeated measures design, 37 raters were required to rate the performance of 4 subordinates who differed from each other in the level of their performance and its consistency. It was hypothesized that performance characteristics would have differential impact on several types of performance accuracy measures. Cronbach's

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

overall accuracy as well as two indices of accuracy based on signal detection theory, labeled classification and behavioral accuracy by Lord (1985), were compared. Results showed that characteristics of ratee performance did affect the accuracy measures differentially.

Keywords:

FD19

# The Impact of Performance Consistency and Performance Level on Alternative Measures of Rater Accuracy

Researchers in the area of performance appraisal have identified a number of factors thought to influence performance ratings. These are: (1) the appraisal instrument; (2) the rater; (3) the ratee; and (4) the performance rating context. Most of the work, thus far, has focused on either the rating instrument or the rater. With respect to the instrument, the major concerns have been increasing the reliability and validity of ratings and reducing rating errors such as halo and leniency (e.g., Latham & Wexley, 1977; Smith & Kendall, 1963). Similarly, the primary outcome of interest when studying rater characteristics has been the reduction of rating errors (Borman, 1979; Cascio & Valenzi, 1977; Taft, 1955).

Recently, it has been argued that the primary goal of performance appraisals should be to obtain ratings that reflect, to the extent possible, the actual behavior of the ratee (Bernardin & Pence, 1980; Borman, 1978). The appropriate criterion for performance appraisal from this perspective is rating accuracy. Given this, research should focus on identifying factors that reduce the ability of raters to make accurate ratings. Unfortunately, while many potential inhibitors of accuracy in performance evaluations have been suggested (Feldman & Hilterman, 1977; Terborg & Ilgen, 1975), empirical research is lacking.

The present research addresses rater accuracy by focusing on one very prominent feature of the rating stimulus--the ratee him or herself. Most of the research on the role of the ratee in performance appraisal has focused on qualities of the ratee that are relatively unchanging over time, such as the ratee's race or gender, and examined their impact on the quality of performance ratings (Hamner, Kim, Baird, & Bigoness, 1974). This approach to studying the ratee is important because static ratee characteristics are likely to be used by raters as cues in forming an initial impression.

In contrast to static information about ratees are dynamic cues which change over time and, thus, result in impression modification. One such dynamic quality is ratee job performance. Although ratee job performance is the criterion against which to judge the validity of ratings, the nature of that performance information may introduce errors into ratings by influencing the way in which the information is processed. There are at least two characteristics of ratee performance that may act to reduce the accuracy of performance appraisals: level (good vs. poor performer) and consistency over time (consistent vs. inconsistent performer). Both characteristics have been addressed in previous research (e.g., DeNisi & Stevens, 1981; Gordon, 1970; Scott & Hamner, 1975) but they have typically been related to such dependent variables as attributions of causality, allocation of organizational rewards and ratings of motivation and ability rather

than to rating accuracy. The present study will examine the effect of ratee performance level and performance consistency on several different measures of rating accuracy.

Ratee Performance Level

Those studies which have compared the impact of known ratee performance along with other variables have found that performance is the best predictor of performance ratings (Bigoness, 1976; Hamner et al., 1974; Leventhal, Perry & Abrami, 1977), but that it accounts for little more than 30% of the variance in ratings (Hamner et al., 1974). Cues other than the actual performance behavior obviously influence the ratings. This suggests that nonperformance-related information can potentially account for a large proportion of the variance in performance ratings and, thus, reduce their validity. In fact, Gordon (1970) found that the level of performance, itself, influenced sensitivity to performance relevant behaviors. He found that, on the average, raters correctly identified 88% of the desirable behaviors exhibited by ratees but only 73% of the undesirable ones and labeled this effect the Differential Accuracy Phenomenon (DAP). One purpose of the present study was to test the DAP.

Ratee Performance Consistency

A second characteristic of performance which may influence performance ratings is the consistency of the performance. Research on performance consistency has dealt with three issues: (1) assessing the ability of raters to make judgments concerning

the consistency of a ratee's performance (Borman, 1983); (2) developing performance appraisal systems that take into account ratee performance inconsistency (Kane & Lawler, 1979); and (3) determining the impact of performance inconsistency on ratings.

Studies in the latter area are most similar to the present research. The few studies conducted (e.g., DeNisi & Stevens, 1981; Scott & Hamner, 1975) suggest that performance consistency does influence performance ratings. In general, variable performance tends to result in more negative ratings. For example, although variable performers were rated as having greater ability, they were given lower ratings of motivation when compared to consistent performers (Scott & Hamner, 1975). Similarly, DeNisi and Stevens (1981) found that among low performers, variable performers received more negative ratings on a composite variable (consisting of ratings of ability and motivation, allocation of organizational rewards and performance attributions) than did stable performers. Both studies found evidence for a recency effect with ascending performance receiving more favorable ratings. Although these studies indicate that ratee performance consistency affects ratings, they do not provide any insight into the effect of consistency on rating accuracy. The present study explicitly examined the effect of performance consistency on several measures of rating accuracy as well as on the rating process in general.

## Hypotheses

Performance Consistency and Sampling. One relevant rating process variable is the amount of information sampled about ratee performance collected by the rater. The cognitive processing view of performance appraisal suggests that there will be an inverse relationship between performance consistency and sampling. According to this view, during a rater's initial exposure to ratees, she or he attempts to form some kind of impression of ratees. This involves placing them into categories that facilitate making sense (Weick, 1979) of their behavior. For consistent performers this categorization process should not be problemmatic since the ratee clearly can be categorized or labeled as either good or poor performer. Furthermore, subsequent behaviors of consistent performers should match the initial categorization and not be questioned. On the other hand, when a ratee's performance is inconsistent, initial categorization becomes more difficult and later observations will fail to fit the category, resulting in controlled re-categorization processes (Feldman, 1981). Consistent with this notion, research indicates that disconfirmed expectations about a stimulus person trigger the perceiver's search for causal information (Pysczynski & Greenberg, 1981; Wong & Weiner, 1981). This suggests our first two hypotheses:

H1: Perceived rating difficulty will be greater for inconsistent performers than for consistent performers.

H2: More sampling of information about ratee performance will occur for inconsistent performers than for consistent performers.

Performance Consistency and Accuracy. The relationship between performance consistency and accuracy is more complex. While it could be argued that the greater ease of categorizing and, thus, evaluating consistent performers should result in greater rating accuracy, it might also be argued that the greater amount of information obtained from sampling for inconsistent performers would result in greater rating accuracy (Favero & Ilgen, 1985; Henemen & Wexley, 1983). The former suggests greater accuracy in evaluating consistent performers while the latter suggests that accuracy will be greater for inconsistent performers.

The two explanations just offered, which appear, at first glance, to be antithetical really are not because each assumes a different conceptualization of rating accuracy. These different conceptualizations of rating accuracy were described by Lord (1985), who used signal detection theory to distinguish between what he called classification accuracy and behavioral accuracy. Classification accuracy is the apparent accuracy that results when raters evaluate people based on general impressions. To the extent that the ratee's actual behavior is consistent with the impression, ratings made on the basis of this impression will appear to be accurate (i.e., classification accuracy will be high). In contrast, behavioral accuracy reflects the ability of raters to identify specific behaviors exhibited by a ratee.

Conceptually these two types of accuracy are not necessarily related to each other and may even covary negatively. For example, classification accuracy is likely to be highest when a rater is able to simplify the rating process by integrating ratee behavior into a single cognitive category and then evaluating the ratee in accordance with the categorization. In this situation, however, behavioral accuracy is likely to be low since integrating ratee behaviors into a cognitive category tends to result in specific behaviors being forgotten. Other examples could be given where the two forms covary positively. Thus, although behavioral accuracy is what is typically meant when the term "accuracy" is used, most common operational definitions of accuracy measure classification accuracy. Specifically, accuracy as measured by the components of accuracy identified by Cronbach (1955) only requires that raters be able to form a general impression of ratees and then evaluate them on the basis of this impression. The knowledge of actual ratee behaviors required for behavioral accuracy is not necessary in order to appear accurate.

When conceptualized as classification accuracy, the primary requirement for accuracy is that raters be able to place ratees into global categories. Classification accuracy should be higher for consistent performers because of the greater ease with which an impression can be formed about these ratees. In addition, sampling may help raters to develop and stabilize an impression of ratees

and, therefore, increase classification accuracy. Thus, we hypothesize that:

H3: There will be a positive relationship between the amount of sampling and classification accuracy.

H4: Rating accuracy as measured by classification-type accuracy measures will be greater for consistent performers than for inconsistent performers.

The reverse is hypothesized for the relationship between behavioral accuracy and performance consistency. There are three reasons for this prediction. The first relates to our hypothesis that inconsistent performance will lead raters to sample more performance information. Presumedly those who observe more behavior should be more likely to recall those behaviors observed and, thus, score higher on behavioral accuracy measures. Second, previous research implies that information inconsistent with expectations is stored in memory in a unique way and, thus, is more likely to be recalled (Graesser, Gordon & Sawyer, 1979; Graesser, Woll, Kowalski & Smith, 1980; Woll & Graesser, 1982). Research in the leadership area is consistent with this notion (e.g., Phillips & Lord, 1982; Phillips, 1984). In addition, Hastie (1980) found that memory for behavior that is inconsistent with general impressions is greater than is memory for behavior consistent with these impressions.

A third reason for the hypothesized greater behavioral accuracy for inconsistent performers is that the greater difficulty

of categorizing behaviors observed from inconsistent ratees should force raters to focus more on specific ratee behaviors when evaluating their performance than on a general impression. This should reduce the tendency for raters to forget ratee behaviors and to attribute to ratees category-consistent behaviors which they did not exhibit. The mechanisms just discussed suggest the following hypotheses:

H5: The amount of sampling of behavioral information will be positively related to behavioral accuracy.

H6: Raters will correctly identify more behaviors for inconsistent performers than for consistent performers.

H7: Raters will be more likely to attribute nonpresent behaviors consistent with the ratee's category to consistent performers than to inconsistent performers.

H8: Rating accuracy as measured using behavioral accuracy will be greater for inconsistent performers than for consistent performers.

Performance Level. Based on the differential accuracy phenomenon identified by Gordon (1970), raters should be more accurate evaluating good performers than poor performers. Since there is no reason to expect a difference between accuracy measures on performance level it is hypothesized that:

H9: Classification and behavioral accuracy will be greater for good performers than for poor performers.

## Method

### Overview

Undergraduates were hired to play the role of an office manager who supervised four secretaries. The four secretaries worked as a team in order to complete work assignments for the faculty members in the department. Four videotapes, one of each secretary, allowed supervisors to observe each secretary at work.

In order to enhance the realism of the setting several conditions described by Ilgen and Favero (1985) as highly desirable for performance appraisal research were incorporated into the study. First, the participants observed the secretaries over time. Second, participants had a variety of different tasks to perform only one of which was evaluating performance. Third, the job of secretary in a department at a university was chosen because it was familiar to the participants and, thus, the social categories used to judge people in this position should be relatively accessible to the subject population (Fiske & Kinder, 1981). Finally, multiple ratees (four secretaries) were evaluated.

### Subjects and Design

A sample of 37 individuals (14 males and 23 females), ranging in age from 17 to 38 years (mean = 22 years) participated in the study. Sample size was based on a power analysis assuming a small effect size and desiring power of .85 (Cohen & Cohen, 1983). Participants were recruited through a newspaper advertisement offering to pay approximately $18 for four hours of participation.

The design was a 2 x 2 analysis of variance with repeated measures on both of the independent variables. The independent variables were performance level (good or poor performer) and performance consistency (consistent and inconsistent). Each participant was exposed to four stimulus persons each representing a combination of the two independent variables (i.e., a consistently good performer, a consistently poor performer, an inconsistently good performer and an inconsistently poor performer).

Development of Stimulus Materials

Construction of Videotapes. Four videotapes were developed, one for each of four secretaries. Each tape contained 17 one to two minute incidents for each secretary with each incident representing some level of performance on one or more of four performance dimensions. The performance dimensions were: (1) job knowledge and skill, (2) organizational ability, (3) dealing with faculty/students, and (4) working cooperatively with other secretaries. Each videotape depicted between 8 and 12 examples of ratee job behavior for each of the four performance dimensions.

To develop behavioral incidents for the videotapes, five secretaries were interviewed and asked to describe incidents of good and poor secretarial performance. One hundred and one incidents were collected from these interviews and were then converted into short descriptions suitable for filming. These incident descriptions were evaluated by six organizational behavior

graduate students on the basis of the dimensions judged to be relevant to the incident and the level of performance represented by the incident on each of those dimensions. These initial ratings were used only to decide which incidents to film for each secretary in order to create the desired manipulations. They were not used in developing the performance standards. Four individuals with secretarial experience were hired to play the roles of the four secretaries in the videotapes. Volunteers filled other roles.

Development of Performance Standards. Ten persons working full time as secretaries rated the videotaped incidents to establish the performance standards against which to judge accuracy. Several procedures were followed to enhance the ability of the secretaries serving as expert raters to provide accurate performance ratings. First, they were given an hour of training on common rating errors (halo, leniency/strictness, central tendency, first impression and contrast effect) and the use of the rating scale. Next, they then practiced rating and discussed their ratings among themselves in order to establish a common frame of reference for each performance dimension (Bernardin & Pence, 1980). In addition, the secretaries were: (1) given a detailed written description of each incident to read prior to each rating session, (2) shown each incident twice before making their rating, and (3) encouraged to take notes.

In making the actual ratings, the expert raters were first asked to indicate which of the four performance dimensions were

represented in the incident. Next, they rated the level of performance on the selected dimension(s) using seven point BARS developed specifically for clerical workers. In some cases, a videotaped incident involved more than one of the four secretaries. When this occurred, all relevant secretaries on the tape were rated.

Using initial criteria of 70% agreement on the dimensions represented by the videotaped incident and a standard deviation of less than 1.25 on the level of the expert ratings, 45 incidents were acceptable, three clearly unacceptable, and 35 were not clearly either acceptable or unacceptable. The raters reconvened to evaluate the 35 incidents for which evaluations were not clear. The second rating of these incidents resulted in 21 which clearly met the criteria and 14 which did so except for one outlier. It was decided to include all 35 in the set of 80 from which the final incidents were chosen as described below.

In order to create the desired performance level and performance consistency manipulations the following criteria were used in selecting specific incidents for each secretary from the pool of 80 incidents: (1) minimize the difference in the average performance level between consistency conditions and maximize this difference between performance level conditions; (2) maximize the variance in performance level across incidents within a dimension for inconsistent performers and minimize this variance for consistent performers; and (3) maximize the variance across

performance dimension means scores for inconsistent performers and minimize this variance for consistent performers. Table 1 presents these data for the incidents that comprised the final four videotapes.

Psychometric Properties of Expert Ratings. In order to be confident that the performance standards developed for each secretary on each performance dimension were indicative of the secretary's actual performance, it was necessary to establish more precisely the extent of agreement between the expert raters on the ratings given to the secretaries. Agreement in two areas was assessed: (1) agreement on the incidents determined to be relevant to each performance dimension, and (2) agreement on the level of performance represented in the incident on the relevant performance dimensions.

The extent of agreement on the incidents judged to be relevant to each dimension was assessed using coefficient alpha. Data were coded "0" (the dimension was not judged to be relevant to the incident) and "1" (the dimension was judged to be relevant to the incident). Alpha coefficients were calculated for each dimension both across incidents for each secretary and across all incidents regardless of secretary. Since the alpha coefficients for each secretary did not differ substantially from the overall alphas, only the alpha coefficients based on all four secretaries are reported. The coefficient alpha was .90 for Job Knowledge and Skill, .96 for Dealing with Professors, .97 for Working

Table 1

Characteristics of Videotapes Used as Stimulus Materials

| Videotape Characteristics | Stimulus Materials | | | |
|---|---|---|---|---|
| | High Performer Consistent | High Performer Inconsistent | Low Performer Consistent | Low Performer Inconsistent |
| 1. Number of behavioral samples | 38 | 40 | 44 | 38 |
| 2. Mean performance level across all behavioral samples | 5.92 | 5.20 | 3.37 | 3.58 |
| 3. Standard deviations across dimension means | .04 | .56 | .70 | .84 |
| 4. Mean standard deviation in performance level across incidents within a dimension | .42 | 1.21 | .81 | 1.36 |

Cooperatively with Other Secretaries, and .91 for Organizational Ability.

Intraclass correlations were used to determine the degree of agreement among the ten expert raters on the overall level of performance of each secretary on each of the four performance dimensions. These were .88 for Job Knowledge, .81 for Dealing with Professors, .87 for Working Cooperatively with Other Secretaries, .93 for Organization of Work, and .91 for the overall ratings.

Procedures

Each person participated in three separate sessions. The first lasted 90 minutes, the second 60, and the final, two hours. Performance ratings were made at the end of the first session and during the final session. Sessions were separated by approximately 10 to 14 days with the first and last sessions 23 to 27 days apart. All sessions were run in a small office on campus. The room was furnished with a table and chair where participants worked on the in-basket tasks. A bookcase served as a partition in the room and behind it was a video recorder and monitor. The monitor was placed so that it could not be seen from the table.

Session 1. When participants arrived for the first session, they read and signed a consent form and filled out an initial questionnaire designed to gather background information. After completing this questionnaire, participants watched a 15 minute videotaped set of instructions which described the study. They were provided with a written copy of the information presented in

the videotape (labeled as the Management Department Office Manual) to read as they listened to the videotape.

The study was described as one that dealt with how managers balanced their time between competing tasks. Participants were told that they would be playing the role of an office manager in the Department of Management at a fictitious university. A brief description of the organization, the job of office manager, and the nature and number of subordinates associated with the position was provided.

Most of the office manager's tasks were presented in an in-basket. The in-basket included such things as filling out a variety of university forms, updating a departmental account, writing several memos, making changes in the course schedule book, scheduling rooms and times for the classes of departmental faculty and making a schedule for the visit of a prospective job candidate In all, 22 tasks were included in the in-basket.

In addition to working on the in-basket, participants were told that they would be evaluating the performance of each of the four secretaries at the end of the first and third sessions and that they could "watch" the secretaries by viewing a portion of a videotape on each secretary. The amount of time they spent viewing the videotapes was up to them after the initial introduction. The evaluation form and its use were then explained to participants.

Participants were informed that they would be evaluated on two major criteria, the quality and quantity of the in-basket work

completed and the accuracy of their performance evaluations of each secretary. As an incentive to perform well, a $25.00 reward was offered for being the highest performer. This was awarded in addition to the $18 paid to everyone for participating in the study.

After completion of the videotaped instructions, the experimenter explained in detail how to operate the video recorder. Then each participant was shown the first five behavioral incidents on the videotape for each secretary. The order in which participants viewed the four secretary videotapes was counterbalanced. Participants were permitted to take notes while watching the videotapes. After watching the videotape of each secretary, the experimenter briefly reviewed the instructions. Participants were then given 30 minutes to complete a performance evaluation on each of the four secretaries and to begin work on the in-basket tasks. Participants were not permitted to view any additional videotape incidents prior to completing the rating form.

Session 2. Session 2 was held approximately 12 days after the first session. When participants arrived for the second session, the instructions, rules, financial incentives, and operation of the equipment were briefly reviewed. Participants were then shown a behavioral incident viewed in the first session to be sure that they could identify each secretary. During the session, the experimenter checked on each participant two times to answer any

questions they might have and to let them know how much time remained.

Session 3. Initial procedures for Session 3 were the same as those for Session 2. Participants were given 50 minutes to work on the in-basket tasks and to watch the videotapes of the secretaries. After 50 minutes, participants were asked to complete a performance evaluation for each of the secretaries and to fill out the final questionnaires. When all questionnaires were completed, participants were debriefed and paid for their participation.

## Measures

Sampling. The amount of information search or sampling done by each rater for each secretary was measured as the total number of behavioral incidents watched for that secretary. At the end of each session, the experimenter recorded the number of incidents watched for each secretary. Prior to beginning the next session the videotapes were set at the point where that person had stopped watching the previous session.

Cronbach's Overall Accuracy. The Cronbach measure of classification accuracy used was overall accuracy. This is measured as a standard sum of squared deviations of the rater's rating on a given dimension from the true rating on that dimension. It was calculated as:

$$OA = \sum_{i=1}^{4} (X_i - S_i)^2$$

where $X_i$ is the rating given to the ratee on dimension i and $S_i$ is the standard on dimension i.

Lord's Accuracy Measures. The information to calculate Lord's (1985) classification and behavioral accuracy measures was derived from a checklist completed by participants during the last session. This checklist consisted of a list of 55 behaviors. Each of the behaviors listed was a major behavior displayed in one of the incidents on the videotape. Sample behavioral statements included: (1) does not get a professor's presentation notes and overheads typed on time, and (2) agrees to stay after regular working hours to finish typing an important paper for a professor in the department. Between 15 and 17 of the behaviors were exhibited by each secretary (there was some overlap since some of the behaviors were exhibited by more than one secretary). For consistent performers, all of the behaviors were consistent with the prototype of either a good or poor performing secretary while for inconsistent performers, approximately half of the behaviors on the checklist were consistent with the prototype.

Participants were asked to read each statement and to indicate which secretary (or secretaries) exhibited that behavior. They were also told to indicate which of the behaviors on the checklist they did not observe. The latter was included as an option because the participants did not watch all of the incidents for each secretary and, thus, would not have seen some of the behaviors on the checklist.

From these responses, it was possible to calculate the prototypical and nonprototypical hit rate and false alarm rate for each of the four secretaries. These were used to determine classification and behavioral accuracy. The prototypical hit rate was the proportion of category-consistent items correctly identified as having been exhibited by the ratee. The prototypical false alarm rate was the proportion of category-consistent nonpresent items (i.e., not exhibited by that secretary) falsely recognized as having been exhibited by the ratee. The nonprototypical hit rate was the proportion of category-inconsistent items correctly attributed to the ratee. Finally, the nonprototypical false alarm rate was the proportion of category-inconsistent nonpresent items incorrectly identified as having been exhibited by the ratee. Note that for consistent performers the nonprototypical hit rate was zero, since, due to the nature of the manipulation, these performers only exhibited behaviors consistent with the category of either good or poor performer.

From these hit rates and false alarm rates, it was then possible to calculate the classification and behavioral accuracy measures. Classification accuracy was calculated as follows:

$$CA = (PHR + PFAR) - (NPHR + NPFAR)$$

where CA is classification accuracy, PHR is the prototypical hit rate, PFAR is the prototypical false alarm rate, NPHR is the nonprototypical hit rate and NPFAR is the nonprototypical false alarm rate. High classification accuracy means that prototypical

behaviors were attributed to the ratees regardless of whether or not they were actually exhibited by them. Behavioral accuracy was calculated using the following formula:

$$BA = (PHR + NPHR) - (PFAR + NPFAR)$$

where BA is behavioral accuracy and the other terms are as defined above. The greater the degree of behavioral accuracy the more able raters are to identify the actual behaviors exhibited by the ratee regardless of the prototypicality of the actual behavior.[1]

## Results

### Manipulation Checks

Two manipulation checks were carried out. First, on the final questionnaire were two questions assessing the perceived consistency of each secretary's performance (average alpha = .68; a separate alpha had to be computed for each secretary) and four items assessing their perceived performance level (average alpha = .76). Each of these scales was used as the dependent variable in a repeated measures analysis of variance.

Results for performance level indicated a highly significant performance level main effect ($F(1,36) = 520.95$, $p < .01$), with the perceived performance level being higher for the high performers than for the low performers ($\bar{x} = 25.03$ vs. $\bar{x} = 11.20$). When the dependent variable was the perceived consistency of the secretary's performance, a significant main effect for consistency was found ($F(1,36) = 75.65$, $p < .01$). Examination of the mean differences revealed that, as expected, the high consistency performers were

perceived as more consistent than the low consistency performers ($\bar{x}$ = 9.70 vs. $\bar{x}$ = 7.30). However, there was also a significant performance level effect on consistency ($F(1,36) = 86.43$, $p < .01$) with high performers being perceived as more consistent than low performers ($\bar{x}$ = 9.93 vs. $\bar{x}$ = 7.07). This effect may have been due to the fact that consistency of performance was perceived positively while inconsistency was perceived negatively and participants tended to attribute positive characteristics to good performers and negative characteristics to poor performers.

Post hoc questioning of each participant provided another check on the manipulations. Ninety-eight percent of the participants correctly identified the performance level of all four secretaries and 77% correctly indicated the degree of consistency for all four. For consistency, the remaining 23% of the participants correctly identified the degree of consistency for two of the four secretaries. There was no apparent pattern of misidentification suggesting that the misidentifications resulted from individual differences in perceptions of the four secretaries rather than from the ineffectiveness of the intended manipulations. In combination, these two manipulation checks provide support for the effectiveness of the performance level and performance consistency manipulations.

Performance Consistency

Cell means and standard deviations for all the dependent variables are reported in Table 2. Marginal means and the results

for the repeated measures analyses of variance used to test the hypotheses are presented in Tables 3 and 4, and will be described more completely below.

Sampling and Rating Difficulty. The first hypothesis stated that raters would find it easier to evaluate consistent performers than inconsistent ones. This hypothesis was tested with a repeated measures analysis of variance using subjects' post task rating of the difficulty of evaluating each secretary as the dependent variable. As predicted, performance consistency significantly affected rating difficulty ($F(1,36) = 43.75$, $p < .01$) such that it was less difficult to evaluate consistent performers than inconsistent ones (see Table 3). It was also hypothesized that the sampling of behavioral information would be greater for inconsistent performers than for consistent performers (Hypothesis 2). This hypothesis was not supported ($F(1,36) = 1.14$, $p > .05$).

Classification Accuracy Measures. The third hypothesis, which predicted that sampling of information would be positively correlated with classification accuracy, received little support for the Cronbach measure of classification accuracy (overall accuracy). None of the correlations between sampling for a particular secretary and the Cronbach classification accuracy measure were significant (the average correlation based on an r to z transformation was -.05). Somewhat different results were found for the Lord measure of classification accuracy. When correlations were computed by collapsing over performance levels within

Table 2

Cell Means and Standard Deviations for Dependent Variables

| Variable[a] | Consistent High Performer | | Inconsistent High Performer | | Consistent Low Performer | | Inconsistent Low Performer | |
|---|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | M | SD | M | SD |
| 1. Rating Difficulty | 1.43 | .65 | 2.39 | .86 | 2.16 | 1.07 | 3.00 | 1.05 |
| 2. Sampling | 7.62 | 1.85 | 8.30 | 2.72 | 9.59 | 2.97 | 9.54 | 2.86 |
| 3. Overall Accuracy | 1.73 | .51 | 2.53 | .84 | 1.45 | .66 | 2.43 | 1.06 |
| 4. Classification Accuracy - Lord | .84 | .34 | .13 | .29 | .67 | .39 | .17 | .31 |
| 5. Behavioral Accuracy | .44 | .28 | .33 | .24 | .28 | .29 | .34 | .24 |
| 6. Hit Rate | .65 | .24 | .50 | .16 | .56 | .23 | .51 | .20 |
| 7. Prototypical False Alarm Rate | .20 | .19 | .16 | .13 | .25 | .13 | .13 | .15 |

[a]N = 37 for all cells.

Table 3

Marginal Means and Analysis of Variance Tables for Rating Difficulty and Sampling

| Variable | Marginal Means: High | Marginal Means: Low | df | MS | F | R-Square |
|---|---|---|---|---|---|---|
| 1. Rating Difficulty | | | | | | |
| Performance Level (L) | 1.91 | 2.58 | 36<br>1 | 1.04<br>16.89 | 16.17[a] | .31 |
| Performance Consistency (C) | 1.79 | 2.69 | 36<br>1 | .72<br>29.43 | 40.65[a] | .53 |
| L x C | -- | -- | 36<br>1 | .57<br>.108 | .19 | 0 |
| 2. Sampling | | | | | | |
| Performance Level (L) | 7.96 | 9.56 | 36<br>1 | 4.97<br>95.68 | 19.23[a] | .35 |
| Performance Consistency (C) | 8.61 | 8.92 | 36<br>1 | 3.14<br>3.57 | 1.14 | .03 |
| L x C | -- | -- | 36<br>1 | 5.38<br>4.92 | .91 | .02 |

[a] $p \leq .01$

consistency conditions, performance consistency appeared to moderate the relationship between sampling and accuracy such that the correlation was nonsignificant for the consistent ratees ($r(73) = -.11$, $p > .05$) but significant and positive for the inconsistent ratees ($r(73) = .30$, $p < .01$).

The fourth hypothesis was that classification accuracy (using both the Cronbach and Lord measures) would be greater for consistent than inconsistent performing ratees. Cronbach's measure of overall accuracy as well as Lord's measure of classification accuracy were used. A repeated measures analysis of variance was done for both of these variables, with very similar results being obtained for each (see Table 4). Specifically, results revealed a significant main effect of performance consistency on both Cronbach's overall accuracy ($F(1,36) = 44.06$, $p < .01$) and Lord's measure of classification accuracy ($F(1,36) = 108.75$, $p < .01$). In both cases there was greater accuracy in evaluating consistent performers. No significant interactions were found for either variable.

Behavioral Accuracy Measure. Hypothesis 5 stated that the amount of sampling would be positively related to behavioral accuracy. There was little support for this hypothesis. A separate correlation was computed between the number of incidents watched for each secretary and the degree of behavioral accuracy in evaluating that secretary. The average correlation (based on an r to z transformation) was nonsignificant ($r = -.09$). In addition,

Table 4

Marginal Means and Analysis of Variance Tables for Accuracy Measures, Hit Rate and Prototypical False Alarm Rate

| Variable[a] | Marginal Means: High | Marginal Means: Low | df | MS | F | R-Square |
|---|---|---|---|---|---|---|
| 1. Overall Accuracy | | | | | | |
| Performance Level (L) | 2.13 | 1.94 | 36 | .73 | | |
| | | | 1 | 1.41 | 1.92 | .05 |
| Performance Consistency (C) | 1.59 | 2.48 | 36 | .66 | | |
| | | | 1 | 29.25 | 44.06[b] | .55 |
| L x C | -- | -- | 36 | .64 | | |
| | | | 1 | .31 | .48 | .01 |
| 2. Classification Accuracy | | | | | | |
| Performance Level (L) | .49 | .42 | 36 | .05 | | |
| | | | 1 | .15 | 3.09 | .08 |
| Performance Consistency (C) | .76 | .15 | 36 | .13 | | |
| | | | 1 | 13.72 | 108.75[b] | .75 |
| L x C | -- | -- | 36 | .13 | | |
| | | | 1 | .42 | 3.22 | .08 |

(table continued)

Table 4 (cont.)

Marginal Means and Analysis of Variance Tables for Accuracy Measures, Hit Rate and Prototypical False Alarm Rate

| Variable[a] | Marginal Means: High | Marginal Means: Low | df | MS | F | R-Square |
|---|---|---|---|---|---|---|
| 3. Behavioral Accuracy | | | | | | |
| Performance Level (L) | .39 | .31 | 36<br>1 | .07<br>.20 | 2.99 | .08 |
| Performance Consistency (C) | .36 | .34 | 36<br>1 | .03<br>.02 | .72 | .02 |
| L x C | -- | -- | 36<br>1 | .03<br>.29 | 9.34[b] | .21 |
| 4. Hit Rate | | | | | | |
| Performance Level (L) | .58 | .54 | 36<br>1 | .03<br>.08 | 2.40 | .06 |
| Performance Consistency (C) | .61 | .51 | 36<br>1 | .03<br>.36 | 13.23[b] | .27 |
| L x C | -- | -- | 36<br>1 | .04<br>.10 | 2.94 | .08 |

(table continued)

Table 4 (cont.)

Marginal Means and Analysis of Variance Tables for Accuracy Measures, Hit Rate and Prototypical False Alarm Rate

| Variable[a] | Marginal Means: High | Marginal Means: Low | df | MS | F | R-Square |
|---|---|---|---|---|---|---|
| 5. Prototypical False Alarm Rate | | | | | | |
| Performance Level (L) | .18 | .19 | 36 | .02 | | |
| | | | 1 | .004 | .28 | 0 |
| Performance Consistency (C) | .23 | .15 | 36 | .01 | | |
| | | | 1 | .24 | 20.64[b] | .36 |
| L x C | -- | -- | 36 | .019 | | |
| | | | 1 | .042 | 2.12 | .06 |

[a] Note that for variable 1, low scores indicate high accuracy while for variables 2 and 3 it is the reverse.

[b] $p \leq .01$

there was no relationship between the variables when the data were collapsed over all secretaries ($r(73) = -.27$, $p > .05$).

The sixth and seventh hypotheses were that raters would correctly identify more behaviors for inconsistent performers than for consistent performers (measured as the hit rate) and that they would be more likely to attribute nonpresent prototypical behaviors to consistent performers (measured as the prototypical false alarm rate). Results of a repeated measures analysis of variance using the hit rate as the dependent variable found a significant performance consistency main effect ($F(1,36) = 13.23$, $p < .01$) with means in the opposite direction of that hypothesized (see Table 4). Specifically, the hit rate was found to be greater for consistent performers than for inconsistent performers. However, a similar analysis of the prototypical false alarm rate found support for the hypothesis. The performance consistency main effect was significant ($F(1,36) = 20.64$, $p < .01$) with the false alarm rate being greater for consistent performers.

The eighth hypothesis was that behavioral accuracy would be greater for inconsistent performers than for consistent performers. This was tested with a repeated measures analysis of variance using Lord's behavioral accuracy as the dependent variable. Results for the relationship between consistency and behavioral accuracy were somewhat supportive of the hypothesis that behavioral accuracy would be greater for inconsistent performers. Although no significant main effects for behavioral accuracy were found, a

significant performance level by performance consistency interaction was obtained ($F(1,36) = 9.34$, $p < .01$). Examination of the cell means indicated that behavioral accuracy was highest for the consistently *good performer* and lowest for the consistently poor performer, while behavioral accuracy for both inconsistent performers was in between (see Table 4). Analyses of simple main effects within performance levels revealed that, among good performers, behavioral accuracy was significantly greater for the consistent performer than the inconsistent performer ($F(1,36) = 9.13$, $p < .01$). Among poor performers, the effect was not significant ($F(1,36) = 2.18$, $p < .05$).

Although the finding that behavioral accuracy was greatest for the consistently good performer was contrary to prediction, it must be tempered by the results of an additional analysis. Because no participants watched all of the incidents available for any of the secretaries, some of the behaviors on the behavioral checklist were not observed by each participant. This resulted in the possibility that participants could attribute behaviors which they did not actually observe to one of the secretaries. Clearly, this is a reduction in behavioral accuracy since it indicates that raters incorrectly identified the behaviors exhibited by the secretaries. Since this could not be incorporated into Lord's determination of behavioral accuracy, it was analyzed separately by counting, for each participant, the number of prototypical behaviors which they

incorrectly attributed to each secretary. This was then used as the dependent variable in a repeated measures analysis of variance.

Results revealed a significant performance consistency main effect ($F(1,36) = 13.32$, $p < .01$) with significantly more unobserved prototypical behaviors being attributed to consistent performers than to inconsistent performers. This finding suggests that the actual level of behavioral accuracy for consistent performers is lower than it appears to be and, thus, provides some support for the hypothesis that behavioral accuracy may be greater for inconsistent performers.

Performance Level

The last hypothesis was that classification and behavioral accuracy would be greater for good performers than for poor performers. Contrary to this hypothesis, the performance level main effect for Cronbach's overall accuracy measure, was found to be nonsignificant ($F(1,36) = 1.92$, $p = .17$). Using Lord's measure of classification accuracy it was found that accuracy tended to be greater for good performers than for poor performers, but the results were only marginally significant ($F(1,36) = 3.10$, $p = .08$). As reported earlier, there was a significant interaction for behavioral accuracy ($F(1,36) = 9.34$, $p < .01$) such that good performers were rated more accurately than poor ones. However, this only occurred for consistent performers (see Table 4).

## Discussion

Previous research has identified several ratee characteristics (e.g., race and sex) that can influence the accuracy of the performance evaluations. The present study examined two potential dynamic sources of inaccuracy in performance appraisals, ratee performance level and performance consistency, and found that performance consistency influenced the way in which raters processed performance information.

Researchers studying cognitive processes have suggested two possible ways in which raters may process performance information (e.g., Nathan & Lord, 1983; Murphy, Carmen, Martin & Garcia, 1982). In one approach, raters integrate behavioral information into general categories or impressions of people. In this case, behavioral specifics tend to be forgotten and general impressions become the basis for subsequent evaluations. In the second approach, it is thought that either actual behavioral observations are recalled or behavioral observations are integrated into behavioral dimensions which are recalled. In either case, the focus is on a more behaviorally-oriented approach to processing and storing information, which should then facilitate accuracy in rating.

The results of this study suggest that one characteristic of ratee performance, its consistency, may influence the approach used to process information. Specifically, for consistent performers, raters were more likely to integrate specific observed behaviors

into general cognitive categories and then use these categories in making evaluations. For inconsistent performers, the more behavioral approach to processing and storing information appears to be used. Several findings from this study combine to support this conclusion: (1) classification accuracy was found to be greater for consistent performers; and (2) raters were more likely to incorrectly attribute both observed and unobserved prototypical behaviors to consistent performers. These findings indicate that raters tended to attribute prototypical behaviors to consistent performing ratees regardless of whether or not they actually exhibited them. Furthermore, raters processed observed behaviors of consistent ratees by integrating them into a general impression which was used as the basis for making performance evaluations.

Performance consistency and the approaches to processing information also affected both measures of rating accuracy--behavioral and classification accuracy. Behavioral accuracy, the correct identification of actual behaviors which were or were not observed, tended to be greater when evaluating inconsistent ratees. This may have been due to the apparent differences between consistent and inconsistent ratees in how information was processed. Apparently, raters processed and stored specific ratee behaviors for inconsistent performing ratees but only a general impression for consistent ratees. The former approach to processing information should result in greater behavioral accuracy

while the latter should lead to enhanced classification accuracy, as was observed.

The finding that the greatest degree of behavioral accuracy occurred for the consistently good performer was contrary to our hypothesis. At first glance, it appears to contradict the notion that for consistent performers, raters process performance information using general categories. Instead, it suggests that raters stored and recalled specific behaviors. However, since consistent performers did not exhibit any nonprototypical behaviors, relying on a general impression of these ratees would actually increase the rater's probability of correctly attributing prototypical behaviors to these ratees, leading to the higher hit rate which we observed and, other things being equal, increasing behavioral accuracy. This would also explain the significant positive correlation observed between classification and behavioral accuracy for the consistent performers ($r(74) = .33$, $p < .01$). On the other hand, for inconsistent performers this correlation was negative ($r(74) = -.29$, $p < .01$) since, in this case, a general impression would tend to reduce behavioral accuracy by increasing the likelihood that raters would forget nonprototypical behaviors.

Several of the hypotheses of this research were based on the premise that the ratee's performance behavior would influence the time spent observing the ratee and, in turn, observation time would affect accuracy. None of the hypotheses involving observation time were supported. There are a number of reasons why observation time

may not have functioned as expected. The most plausible to us is that the repeated measures nature of the design may have led participants to believe that they ought to sample all ratees approximately equally. Such a belief would have removed differences among ratees and, thus, the hypothesized differences in observation time.

This study also found some support for the Differential Accuracy Phenomenon identified by Gordon (1970) since, among consistent performers, behavioral accuracy was greater for the good performer than for the poor ones. However, there was little difference between good and poor performers who were inconsistent. Raters appeared to find it easier to evaluate good performers, perhaps because they had a better understanding of what constituted good performance than poor performance. Specifically, when someone exhibited an undesirable behavior it may have been difficult for raters to determine how ineffective that behavior was unless it was extremely ineffective. Good behavior, on the other hand, may have been less ambiguous and, therefore, easier to identify. The reason the effect was only found for consistent performers is not clear.

Implications

The results of this study have several implications. First, the distinction between classification and behavioral accuracy is important as is the empirical demonstration of the differences in the way these accuracy measures function. Consistent with Lord (1985), we have argued that although accuracy, in a conceptual

sense, really implies behavioral accuracy, common operationalizations of accuracy (such as Cronbach's) only tap classification accuracy. We also agree with Lord (1985) that researchers in the area of performance appraisal looking at rating accuracy, particularly those studying cognitive processes, should begin to use a more behavioral approach to assessing accuracy. Although this is a more rigorous accuracy criterion than that typically assessed, we believe it is more conceptually correct and, thus, avoids the problem of rater's appearing to be accurate with Cronbach's measures of classification accuracy when, in fact, in a true behavioral sense, they are not.

While a behavioral accuracy criterion also has practical relevance for such functions as providing developmental feedback, in other situations, classification accuracy may be all that is required of raters. For example, when raters have to select a subordinate to receive an award or determine which subordinates should be given the largest or smallest pay increases, all that is necessary is that raters be able to assess, in a general way, the overall performance level of the ratee. Given this, we suggest that classification accuracy (assessed by either the Cronbach or Lord measures) can, perhaps, best be seen as a necessary but not sufficient condition for true behavioral accuracy.

The data from this study also suggest the need to identify factors that might increase the tendency of raters to rely on a general impression in evaluating performance rather than on

specific ratee behaviors since this tends to reduce behavioral accuracy. The present study suggested one such factor, the consistency of a ratee's performance, but other factors, such as the race and sex of the ratee, might also result in a similar tendency. This could account for the occurrence of inaccuracy in ratings. From a practical point of view, this study supports the suggestions of previous researchers (e.g., Bernardin & Pence, 1980) on the importance of training raters to focus on ratee behaviors, using such things as behavioral diaries.

The study also suggests that raters need to learn the distinction between prototypical and nonprototypical behaviors and be aware of the common tendency for false positive prototypical behavior errors. For example, when observing multiple ratees, the rater may recall observing a particular behavior but attribute that behavior to the ratee for which the behavior is most prototypical. One potential outcome of this tendency would be for the ratings of consistent performers to be more extreme. In other words, good performers would generally be rated higher than they deserve while poor performers would tend to be rated lower than they ought to be. If raters can be taught to eliminate this type of error, perhaps through procedures such as reality monitoring (Johnson & Raye, 1981), the accuracy of evaluations might be increased. Since some evidence suggests that observational accuracy is positively related to rating accuracy (Murphy, Garcia, Kerkar, Martin & Balzer, 1982)

future research should examine ways to increase behavioral observation accuracy.

## References

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Borman, W. C. (1978). Exploring the upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 63, 135-144.

Borman, W. C. (1979). Individual differences correlates of accuracy in evaluating others' performance effectiveness. Applied Psychological Measurement, 3, 103-115.

Borman, W. C. (1983). Implications of personality theory and research for the rating of work performance in organizations. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory. Hillsdale, NJ: Erlbaum.

Bigoness, N. J. (1976). Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. Journal of Applied Psychology, 61, 80-84.

Cascio, W. F., & Valenzi, E. R. (1977). Behaviorally anchored rating scales: Effects of education and job experience of raters and ratees. Journal of Applied Psychology, 62, 278-282.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/ correlation analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." Psychological Bulletin, 52, 177-193.

DeNisi, A. S., & Stevens, G. E. (1981). Profiles of performance, performance evaluations, and personnel decisions. Academy of Management Journal, 24, 592-602.

Favero, J. L., & Ilgen, D. R. (1985). The effects of ratee characteristics on rater performance appraisal behavior (Tech. Rep. No. 83-5). Office of Naval Research.

Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

Feldman, J. M., & Hilterman, R. J. (1977). Sources of bias in performance evaluation: Two experiments. International Journal of Intercultural Relations, 1, 35-57.

Fiske, S. T., & Kinder, D. R. (1981). Involvement expertise, and schema use: Evidence from political cognition. In N. Cantor & J. Kihlstrom (Eds.), Personality, cognition and social interaction. Hillsdale, NJ: Erlbaum.

Gordon, M. E. (1970). The effect of the correctness of the behavior observed on the accuracy of ratings. Organizational Behavior and Human Performance, 5, 366-377.

Graesser, A. C., Gordon, S. E., & Sawyer, J. D. (1979). Recognition memory for typical and atypical actions in scripted activities: A test of a script pointer + tag hypothesis. Journal of Verbal Learning and Verbal Behavior, 18, 319-332.

Graesser, A. C., Woll, S. B., Kowalski, D. J., & Smith, D. A. (1980). Memory for typical and atypical actions in scripted activities. Journal of Experimental Psychology: Human Learning and Memory, 6, 503-515.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, N. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work sampling task. Journal of Applied Psychology, 59, 705-711.

Hastie, R. (1980). Memory for behavioral information that confirms or contradicts a personality impression. In R. Hastie, T. M. Ostrom, E. B. Ebbesen, R. S. Wyer, D. L. Hamilton, & D. E. Carlston (Eds), Person memory: The cognitive basis of social perception. Hillsdale, NJ: Erlbaum.

Henemen, R. L., & Wexley, K. N. (1983). The effects of time delay in rating and amount of information observed on performance rating accuracy. Academy of Management Journal, 26, 677-686.

Johnson, M. K., & Raye, C. L. (1982). Reality monitoring. Psychological Review, 88, 67-85.

Kane, J. S., & Lawler, E. E. (1979). Performance appraisal effectiveness: Its assessment and determinants. In B. M. Staw (Ed), Research in organizational behavior Vol. 1. Greenwich, CT: JAI Press.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. Personnel Psychology, 30, 255-268.

Leventhal, L., Perry, R. P., & Abrami, P. C. (1977). Effect of lecturer quality and student perception of lecturer's experience on teacher ratings. Journal of Educational Psychology, 69, 360-374.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. Journal of Applied Psychology, 70, 66-71.

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? Journal of Applied Psychology, 67, 562-567.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluation performance. Journal of Applied Psychology, 67, 320-325.

Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. Journal of Applied Psychology, 68, 102-114.

Phillips, J. S. (1984). The accuracy of leadership ratings: A cognitive categorization perspective. Organizational Behavior and Human Performance, 33, 125-138.

Phillips, J. S., & Lord, R. G. (1982). Schematic information processing and perceptions of leadership in problem-solving groups. Journal of Applied Psychology, 67, 486-492.

Pyszczynski, T. A., & Greenberg, J. (1981). The role of disconfirmed expectations in the investigation of attributional processing. Journal of Personality and Social Psychology, 40, 32-38.

Scott, W. E., & Hamner, W. C. (1975). The influence of variations in performance profiles on the performance evaluation process: An examination of the validity of the criterion. Organizational Behavior and Human Performance, 14, 360-370.

Smith, P., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Taft, R. (1955). The ability to judge people. Psychological Bulletin, 52, 1-23.

Terborg, J. R., & Ilgen, D. R. (1975). A theoretical approach to sex discrimination in traditionally masculine occupations. Organizational Behavior and Human Performance, 13, 352-376.

Weick, K. E. (1979). The social psychology of organizing. Reading, MA: Addison-Wesley.

Woll, S. B., & Graesser, A. C. (1982). Memory discrimination for information typical or atypical of person schemata. Social Cognition, 1, 287-310.

Wong, P. T. P., & Weiner, B. (1981). When people ask "why" questions, and the heuristics of attributional search. Journal of Personality and Social Psychology, 40, 650-663.

## Footnote

[1]A number of other variables not directly related to the major hypotheses of this study were assessed on the final questionnaire but were not reported here.

## LIST 1 MANDATORY*

Defense Technical Information Center (12)
ATTN: DTIC DDA-2
Selection & Preliminary Cataloging Section
Cameron Station
Alexandria, VA 22314

Library of Congress
Science and Technology Division
Washington, DC 20540

Office of Naval Research (3)
Code 4420E
800 N. Quincy Street
Arlington, VA 22217

Naval Research Laboratory (6)
Code 2627
Washington, DC 20375

Office of Naval Research
Director, Technology Programs
Code 200
800 N. Quincy Street
Arlington, VA 22217

## LIST 2 ONR FIELD

Psychologist
Office of Naval Research
Detachment, Pasadena
1030 East Green Street
Pasadena, CA 91106

## LIST 3 OPNAV

Deputy Chief of Naval Operations
(Manpower, Personnel & Training)
Head, Research, Development, and
Studies Branch (OP-115)
1812 Arlington Annex
Washington, DC 20350

Director
Civilian Personnel Division (OP-14)
Department of the Navy
1803 Arlington Annex
Washington, DC 20350

Deputy Chief of Naval Operations
(Manpower, Personnel, & Training)
Director, Human Resource Management
Plans & Policy Branch (OP-150)
Department of the Navy
Washington, DC 20350

## LIST 4 NAVMAT & NPRDC

Program Administrator for Manpower,
Personnel, and Training
MAT-0722
800 N. Quincy Street
Arlington, VA 22217

Naval Material Command
Management Training Center
NAVMAT 09M32
Jefferson Plaza, Bldg #2, Rm 150
1421 Jefferson Davis Highway
Arlington, VA 20360

Naval Material Command
Director, Productivity Management
Office
MAT-00K
Crystal Plaza #5, Rm 632
Washington, DC 20360

Naval Personnel R&D Center (4)
Technical Director
Director, Manpower & Personnel
Laboratory, Code 06
Director, System Laboratory, Code 07
Director, Future Technology, Code 41
San Diego, CA 92152

*Number in parentheses is the number of copies to be sent.

Navy Personnel R&D Center
Washington Liaison Office
Ballston Tower #3, Rm 93
Arlington, VA 22217

LIST 5 BUMED

NONE

LIST 6
NAVAL ACADEMY AND NAVAL POSTGRADUATE SCHOOL

Naval Postgraduate School (3)
ATTN: Chairman, Dept. of
Administrative Science
Department of Administrative Sciences
Monterey, CA 93940

U.S. Naval Academy
ATTN: Chairman
Department of Leadership & Law
Stop 7-B
Annapolis, MD 21402

LIST 7 HRM

Officer in Charge
Human Resource Management Division
Naval Air Station
Mayport, FL 32228

Human Resource Management School
Naval Air Station Memphis (96)
Millington, TN 38054

Commanding Officer
Human Resource Management School
Naval Air Station Memphis
Millington, TN 38054

LIST 8 NAVY MISCELLANEOUS

Naval Military Personnel Command (2)
HRM Department (NMPC-6)
Washington, DC 20350

LIST 9 USMC

Headquarters, U.S. Marine Corps
ATTN: Scientific Adviser,
Code RD-1
Washington, DC 20380

LIST 10 OTHER FEDERAL GOVERNMENT

Dr. Brian Usilaner
GAO
Washington, DC 20548

Social and Developmental Psychology
Program
National Science Foundation
Washington, DC 20550

Office of Personnel Management
Office of Planning and Evaluation
Research Management Division
1900 E. Street, NW
Washington, DC 20415

LIST 11 ARMY

Technical Director (3)
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Head, Department of Behavior
Science and Leadership
U.S. Military Academy
New York 10996

LIST 12 AIR FORCE

Air University Library
LSE 76-443
Maxwell AFB, AL 36112

Head, Department of Behavioral
Science and Leadership
U.S. Air Force Academy
Colorado 80840

LIST 13 MISCELLANEOUS

Dr. Eduardo Salas
Human Factors Division
Code 712
Navy Training Systems Center
Department of the Navy
Orlando, FL 32813-7100

LIST 14 CURRENT CONTRACTORS

Dr. Janet L. Barnes-Farrell
Department of Psychology U-20
University of Connecticut
406 Cross Campus Road
Storrs, CT 06268

Jeanne M. Brett
Northwestern University
Graduate School of Management
2001 Sheridan Road
Evanston, IL 60201

Dr. Terry Connolly
Georgia Institute of Technology
School of Industrial & Systems
Engineering
Atlanta, GA 30332

Dr. Richard Daft
Texas A&M University
Department of Management
College Station, TX 77843

Dr. Randy Dunham
University of Wisconsin
Graduate School of Business
Madison, WI 53706

Dr. Lawrence R. James
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332

Dr. J. Richard Hackman
School of Organization & Management
Box 1A
Yale University
New Haven, CT 06520

Dr. Frank J. Landy
Department of Psychology
Pennsylvania State University
450 Moore Bldg.
University Park, PA 16802

Dr. Bibb Latane
University of North Carolina
at Chapel Hill
Manning Hall 026A
Chapel Hill, NC 27514

Dr. Edward E. Lawler III
Graduate School of Business
University of Southern California
Los Angeles, CA 90007

END

DTIC

6-86